How to Play in Infinite MDPs (Invited Talk)

How to Play in Infinite MDPs

Stefan Kiefer
Department of Computer Science, University of Oxford, United Kingdom

Richard Mayr
School of Informatics, University of Edinburgh, United Kingdom

Mahsa Shirmohammadi
CNRS & IRIF, Université de Paris, France

Patrick Totzke
Department of Computer Science, University of Liverpool, United Kingdom

Dominik Wojtczak
Department of Computer Science, University of Liverpool, United Kingdom

Abstract

Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and nondeterministic behavior. For MDPs with finite state space it is known that for a wide range of objectives there exist optimal strategies that are memoryless and deterministic. In contrast, if the state space is infinite, optimal strategies may not exist, and optimal or ε-optimal strategies may require (possibly infinite) memory. In this paper we consider qualitative objectives: reachability, safety, (co-)Büchi, and other parity objectives. We aim at giving an introduction to a collection of techniques that allow for the construction of strategies with little or no memory in countably infinite MDPs.

2012 ACM Subject Classification Theory of computation → Random walks and Markov chains; Mathematics of computing → Probability and statistics

Keywords and phrases Markov decision processes

Digital Object Identifier 10.4230/LIPIcs.ICALP.2020.3

Category Invited Talk

Funding *Stefan Kiefer*: Supported by a Royal Society Research Fellowship.

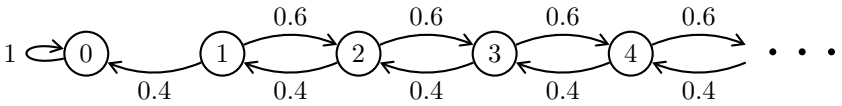
1 Introduction

Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and controlled behavior [13]. MDPs play a prominent role in numerous domains, including artificial intelligence and machine learning [16, 15], control theory [3, 1], operations research and finance [4, 14], and formal verification [8, 2]. In an MDP, the system starts in the initial state and makes a sequence of transitions between states. Depending on the type of the current state, either the controller gets to choose an enabled transition (or a distribution over transitions), or the next transition is chosen randomly according to a defined distribution. By fixing a strategy for the controller, one obtains a probability space of runs of the MDP. The goal of the controller is to maximize the probability of a given objective (some set of desired runs), or, more generally, to optimize the expected value of a random variable (some real-valued function on runs).

The type of strategy needed to satisfy an objective optimally (or ε-optimally) is called the *strategy complexity* of the objective. There are different types of strategies, depending on whether one can take the whole history of the run into account (history-dependent; (H)), or whether one is limited to a finite amount of memory (finite memory; (F)) or whether decisions are based only on the current state (memoryless; (M)). Moreover, the strategy

type depends on whether the controller can randomize (R) or is limited to deterministic choices (D). The simplest type MD refers to memoryless deterministic strategies. *Markov strategies* are strategies that base their decisions only on the current state and the number of steps in the history or the run. Thus they use infinite memory, but only in a very restricted form by maintaining an unbounded step counter.

For finite MDPs, there exist optimal MD-strategies for many (but not all) objectives [5, 6, 7, 13], but the picture is more complex for countably infinite MDPs [11, 12, 13]. For example, given some objective, consider the set of all states for which there exists a strategy that achieves the objective with positive probability. If the MDP is finite then this set is finite and thus there exists some minimal nonzero value, which can often be exploited for the construction of an optimal strategy. These methods do not carry over to infinite MDPs. Here it is possible, even for reachability objectives, that every state has a strategy that achieves the objective with positive probability, but no state, except the target itself, can achieve it almost surely. Such phenomena appear already in infinite-state Markov chains like the classic gambler's ruin problem with unfair coin tosses in the player's favor (0.6 win, 0.4 lose):



The probability of ruin is always positive, but less than 1 in every state except the ruin state itself; cf. [9, Chapter 14]. Another difference to finite MDPs is that optimal strategies need not exist, even for qualitative objectives like reachability or parity. Even if there is a sequence of strategies whose success probabilities converge to 1, there may not exist a strategy with success probability equal to 1. This motivates the investigation of ε-*optimal* strategies, which are those strategies such that no other strategy has a success probability that is more than ε higher.

In this paper we restrict ourselves to MDPs with *countable* state space. Certain theorems such as Theorem 3 are known to be false for MDPs with uncountably many states, see [12]. Uncountable MDPs in general have been studied less, and the underlying measure theory is more complicated.

We aim at providing an introduction to a toolkit for the construction of memoryless or "low-memory" optimal and ε-optimal strategies for certain qualitative objectives like reachability and safety. We will illustrate that these techniques can be combined to construct strategies for more general objectives.

2 Preliminaries

A *probability distribution* over a countable set S is a function $f : S \to [0,1]$ with $\sum_{s \in S} f(s) = 1$. We write $\mathcal{D}(S)$ for the set of all probability distributions over S.

Markov decision processes. In this paper we study Markov decision processes (MDPs) over *countably infinite* state spaces. Formally, an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ consists of a countable set S of *states*, which is partitioned into a set $S_\Box$ of *controlled states* and a set $S_\bigcirc$ of *random states*, a *transition relation* $\longrightarrow \subseteq S \times S$, and a *probability function* $P : S_\bigcirc \to \mathcal{D}(S)$. If $(s, s') \in \longrightarrow$, we call s' a *successor* of s. We assume that every state has at least one successor. The probability function P assigns to each random state $s \in S_\bigcirc$ a probability distribution $P(s)$ over its set of successor states. A *sink* is a subset $T \subseteq S$ closed under the $\longrightarrow$ relation. An MDP is *acyclic* if the underlying directed graph $(S, \longrightarrow)$ is acyclic. It is *finitely branching* if every state has finitely many successors and *infinitely branching* otherwise. An MDP without controlled states ($S_\Box = \emptyset$) is a *Markov chain*.

Strategies and Probability Measures. A *run* ρ is an infinite sequence $s_0 s_1 \cdots$ of states such that $(s_i, s_{i+1}) \in \longrightarrow$ for all $i \in \mathbb{N}$. A *partial run* is a finite prefix of a run. A *strategy* is a function $\sigma : S^* S_\Box \to \mathcal{D}(S)$ that assigns to partial runs $\rho s \in S^* S_\Box$ a distribution over the successors of s.

A strategy σ and an initial state $s_0 \in S$ induce a standard probability measure on sets of infinite runs, see, e.g., [10]. Such measurable sets of infinite runs are called *events* or *objectives*. We write $\mathbb{P}^\sigma_{\mathcal{M},s_0}(E)$ for the probability of an event $E \subseteq s_0 S^\omega$ of runs starting from s_0. We may drop the subscript $\mathcal{M}$ when it is understood.

Objectives. Given a set $T \subseteq S$ of states, the *reachability* objective $\texttt{Reach}(T)$ is the set of runs that visit T at least once; and the *safety objective* $\texttt{Safety}(T)$ is the set of runs that never visit T. *Parity* objectives are defined via a *color function* $Col : S \to \mathbb{N}$ with finite range. The corresponding parity objective is the set of runs such that the largest color that occurs infinitely often along the run is even. We call a parity objective a $\mathcal{C}$-parity objective if $Col(S) \subseteq \mathcal{C}$, i.e., the set of colors is restricted to $\mathcal{C}$. *Büchi* and *co-Büchi* objectives are common names for $\{1,2\}$- and $\{0,1\}$-parity objectives, respectively.

Optimal and ε-optimal Strategies. Given an objective E, the *value* of state s in an MDP $\mathcal{M}$, denoted by $\texttt{val}_{\mathcal{M},s}(E)$, is the supremum probability of achieving E, i.e., $\texttt{val}_{\mathcal{M},s}(E) \stackrel{\text{def}}{=} \sup_{\sigma\in\Sigma} \mathbb{P}^\sigma_{\mathcal{M},s}(E)$ where Σ is the set of all strategies. We may drop the subscript $\mathcal{M}$ when it is understood. For $\varepsilon > 0$ and a state $s \in S$, we say that a strategy is *ε-optimal* if $\mathbb{P}^\sigma_{\mathcal{M},s}(E) \geq \texttt{val}_{\mathcal{M},s}(E) - \varepsilon$. A 0-optimal strategy is called *optimal*. An optimal strategy is *almost surely winning* if $\texttt{val}_{\mathcal{M},s}(E) = 1$.

Strategy Classes. Strategies $\sigma : S^* S_\Box \to \mathcal{D}(S)$ are in general *randomized* (R) in the sense that they take values in $\mathcal{D}(S)$. A strategy σ is *deterministic* (D) if $\sigma(\rho)$ is a Dirac distribution for all partial runs $\rho \in S^* S_\Box$. A strategy is called *memoryless* (M) or *positional* if it only depends on the current state; i.e., a memoryless strategy can be given by a function $\sigma : S_\Box \to \mathcal{D}(S)$. Thus, the simplest strategies are MD strategies, which are both memoryless and deterministic (and thus can be given by a function $\sigma : S_\Box \to S$).

We also consider strategies with memory, but we do not formalize this here. After each transition such a strategy updates its memory mode depending on the taken transition and the previous memory mode. To choose a successor of a controlled state, the strategy can use its memory and the current state but not the partial run that led to the current state. Every strategy can be viewed as a strategy with memory (by using partial runs as memory modes).

For example, k-bit strategies use (at most) k bits of memory; they have (at most) 2^k memory modes. *Markov* strategies use infinite memory but only as a step counter; such strategies depend only on the current state and the number of steps taken so far. A *k-bit Markov* strategy can use both k bits and an (unbounded) step counter.

3 Constructing MD Strategies

In this section we illustrate some techniques to construct MD strategies. We mostly focus on reachability objectives, which we use as a running example. But many ideas apply also to other objectives.

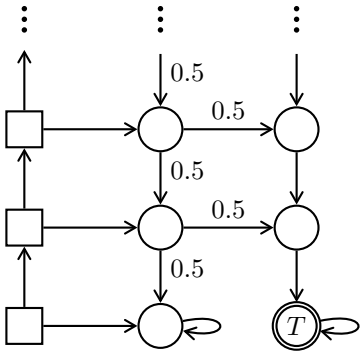


Figure 1 Optimal strategy for reachability may not exist. Here, as in all pictures, we depict controlled states as squares, and random states as circles.

3.1 Reduction to a Finite Case

Suppose we have a (countable) MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$, and we are interested in the reachability objective $\texttt{Reach}(T)$, i.e., we would like to reach a target set $T \subseteq S$.

If S is finite, the situation is as good as it could be:

▶ **Lemma 1** (optimal MD strategies in finite MDPs). *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and consider the reachability objective $\texttt{Reach}(T)$ for some $T \subseteq S$. If S is finite then there exists a single MD strategy σ that is optimal for every start state s at the same time; formally, $\mathbb{P}_s^\sigma(\texttt{Reach}(T)) = \texttt{val}_s(\texttt{Reach}(T))$ for all $s \in S$.*

Spelled out, Lemma 1 says that, if the state space is finite, we can fix for each controlled state an outgoing transition in a way that maximizes the probability of reaching the target for every state. The assumption of a finite state space is so powerful that Lemma 1 generalizes from reachability to parity objectives. In fact, even in finite parity *games*, where some controlled states are controlled by a player who wants to maximize the probability of achieving the parity objective, and the remaining controlled states are controlled by a player who wants to *minimize* the probability of achieving the parity objective, both players have optimal MD strategies for all states [17, Theorem 1].

Lemma 1 does not hold without the assumption of S being finite, as optimal strategies may not exist, let alone optimal MD strategies. Indeed, consider the MDP in Figure 1. The controlled states are those in the leftmost column. Suppose you start in the bottom-left state and you would like to reach the target T consisting only of the bottom-right state. The higher you climb the ladder of states in the left column, the bigger you can make the probability to reach T, but eventually you have to turn right and hope for the best. We have $\texttt{val}_s(\texttt{Reach}(T)) = 1$ for all controlled states s, i.e., we can get the probability of reaching T arbitrarily close to 1. But we cannot make that probability equal to 1: there is no strategy that reaches T with probability 1.

Recall that a strategy σ is called ε-optimal for s if $\mathbb{P}_s^\sigma(\texttt{Reach}(T)) \geq \texttt{val}_s(\texttt{Reach}(T)) - \varepsilon$. The following lemma says that every state has ε-optimal *MD* strategies:

▶ **Lemma 2** (non-uniform ε-optimal MD strategies). *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and consider the reachability objective $\texttt{Reach}(T)$ for some $T \subseteq S$. For every $\varepsilon > 0$ and every $s \in S$ there exists an MD strategy σ that is ε-optimal for s; formally, $\mathbb{P}_s^\sigma(\texttt{Reach}(T)) \geq \texttt{val}_s(\texttt{Reach}(T)) - \varepsilon$.*

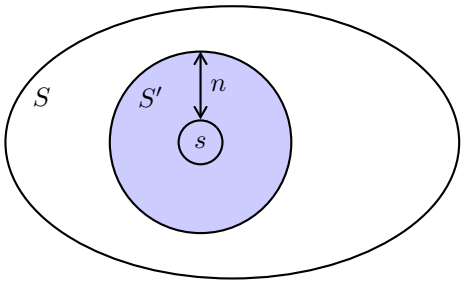


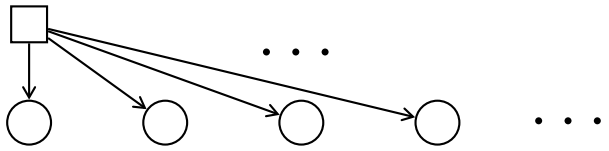
Figure 2 The set S' consists of those states that are reachable from s within at most n steps. Due to finite branching, S' is finite. For n large enough, S' probably suffices to reach the target.

Proof. First assume that $\mathcal{M}$ is finitely branching. Fix $s \in S$ and $\varepsilon > 0$, and let τ be an arbitrary (i.e., not necessarily MD) strategy with $\mathbb{P}_s^\tau(\mathtt{Reach}(T)) \geq \mathtt{val}_s(\mathtt{Reach}(T)) - \frac{\varepsilon}{2}$. Define a modified reachability objective $\mathtt{Reach}_i(T)$ which means reaching T in exactly i steps. Then we have:

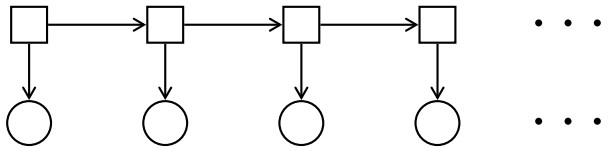
$$\sum_{i=0}^{\infty} \mathbb{P}_s^\tau(\mathtt{Reach}_i(T)) \;=\; \mathbb{P}_s^\tau(\mathtt{Reach}(T)) \;\geq\; \mathtt{val}_s(\mathtt{Reach}(T)) - \frac{\varepsilon}{2}$$

It follows that we can pick a number n large enough so that $\sum_{i=0}^{n} \mathbb{P}_s^\tau(\mathtt{Reach}_i(T)) \geq \mathtt{val}_s(\mathtt{Reach}(T)) - \varepsilon$. From state s, in at most n steps, strategy τ can only use a finite subset $S' \subseteq S$, as $\mathcal{M}$ is finitely branching; see Figure 2. That means, τ manages to reach T with probability at least $\mathtt{val}_s(\mathtt{Reach}(T)) - \varepsilon$ even when it is restricted to the sub-MDP with state space S' (think of leaving S' as losing). But by Lemma 1 this finite sub-MDP has an optimal MD strategy σ. We may extend the definition of σ outside of S' in an arbitrary way. Then, in $\mathcal{M}$, we have $\mathbb{P}_s^\sigma(\mathtt{Reach}(T)) \geq \mathtt{val}_s(\mathtt{Reach}(T)) - \varepsilon$, as desired.

The assumption that $\mathcal{M}$ is finitely branching can be satisfied using a simple construction: replace every infinitely branching controlled state



by



A similar construction works for random states. Then, construct an MD strategy, as above, from an arbitrary ε-optimal strategy in the new finitely branching MDP. Any MD strategy can be transferred back to the original infinitely branching MDP. ◀

3.2 Ornstein's Plastering Technique

Ornstein [12] proved in 1969 a uniform version of Lemma 2. That is, for every $\varepsilon > 0$ there exists a single MD strategy that is ε-optimal for all states:

▶ **Theorem 3** ([12], uniform ε-optimal MD strategies). *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and consider the reachability objective $\mathtt{Reach}(T)$ for some $T \subseteq S$. For every $\varepsilon > 0$ there exists a single MD strategy σ that is ε-optimal for every start state s at the same time; formally, $\mathbb{P}^\sigma_s(\mathtt{Reach}(T)) \geq \mathtt{val}_s(\mathtt{Reach}(T)) - \varepsilon$ for all $s \in S$.*

Ornstein actually proved a stronger statement with multiplicative instead of additive error; i.e., Theorem 3 also holds with $\mathtt{val}_s(\mathtt{Reach}(T)) - \varepsilon$ replaced by $\mathtt{val}_s(\mathtt{Reach}(T)) \cdot (1 - \varepsilon)$.

Proof of Theorem 3. We follow Ornstein's proof [12]. Without loss of generality, we assume that T is a sink. Recall that an MD strategy σ can be viewed as a function $\sigma : S_\Box \to S$ such that for all $s \in S_\Box$, the state $\sigma(s)$ is a successor state of s. Starting from the original MDP $\mathcal{M}$ we successively *fix* more and more controlled states, by which we mean select an outgoing transition and remove all others. While this is in general an infinite (but countable) process, it defines an MD strategy in the limit. Visually, we "plaster" the whole state space by the fixings.

Put the states in some order, i.e., $s_1, s_2, \ldots$ with $S = \{s_1, s_2, \ldots\}$. The plastering proceeds in *rounds*, one round for every state. Let $\mathcal{M}_i$ be the MDP obtained from $\mathcal{M}$ after the fixings of the first $i-1$ rounds (with $\mathcal{M}_1 = \mathcal{M}$). In round i we fix controlled states in such a way that

(A) the probability, starting from s_i, of reaching the target T using only random and *fixed* controlled states is not much less than the value $\mathtt{val}_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T))$; and

(B) for all states s, the value $\mathtt{val}_{\mathcal{M}_{i+1}, s}(\mathtt{Reach}(T))$ is almost as high as $\mathtt{val}_{\mathcal{M}_i, s}(\mathtt{Reach}(T))$.

The purpose of goal (A) is to guarantee good progress towards the target when starting from s_i. The purpose of goal (B) is to avoid fixings that would cause damage to the values of other states.

Now we describe round i. Consider the MDP $\mathcal{M}_i$ after the fixings from the first $i-1$ rounds, and let $\varepsilon_i > 0$. Recall that we wish to fix a part of the state space so that s_i has a high probability of reaching T using only random and fixed controlled states. By Lemma 2 there is an MD strategy σ such that $\mathbb{P}^\sigma_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T)) \geq \mathtt{val}_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T)) - \varepsilon_i^2$. Fixing σ everywhere would accomplish goal (A), but potentially compromise goal (B). So instead we are going to fix σ only for states where σ does well: define

$$G \stackrel{\text{def}}{=} \{s \in S \mid \mathbb{P}^\sigma_{\mathcal{M}_i, s}(\mathtt{Reach}(T)) \geq \mathtt{val}_{\mathcal{M}_i, s}(\mathtt{Reach}(T)) - \varepsilon_i\}$$

and obtain $\mathcal{M}_{i+1}$ from $\mathcal{M}_i$ by fixing σ on G. (Note that σ does not "contradict" earlier fixings, because in the MDP $\mathcal{M}_i$ the previously fixed states have only one outgoing transition left.) See Figure 3 for an illustration.

We have to check that with this fixing we accomplish the two goals above. Indeed, we accomplish goal (A): by its definition strategy σ is ε_i^2-optimal from s_i, so the probability of ever entering $S \setminus G$ (where σ is less than ε_i-optimal) cannot be large:

$$\mathbb{P}^\sigma_{\mathcal{M}_i, s_i}(\mathtt{Reach}(S \setminus G)) \;\leq\; \varepsilon_i \tag{1}$$

In slightly more detail, this inequality holds because the probability that the ε_i^2-optimal strategy σ enters a state whose value is underachieved by σ by at least ε_i can be at most ε_i. We give a detailed proof of (1) in Section 5.1. It follows from the ε_i^2-optimality of σ and from (1) that we have $\mathbb{P}^\sigma_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T) \wedge \neg\mathtt{Reach}(S \setminus G)) \geq \mathtt{val}_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T)) - \varepsilon_i - \varepsilon_i^2$. So in $\mathcal{M}_{i+1}$ we obtain for *all* strategies σ':

$$\mathbb{P}^{\sigma'}_{\mathcal{M}_{i+1}, s_i}(\mathtt{Reach}(T)) \;\geq\; \mathtt{val}_{\mathcal{M}_i, s_i}(\mathtt{Reach}(T)) - \varepsilon_i - \varepsilon_i^2 \tag{2}$$

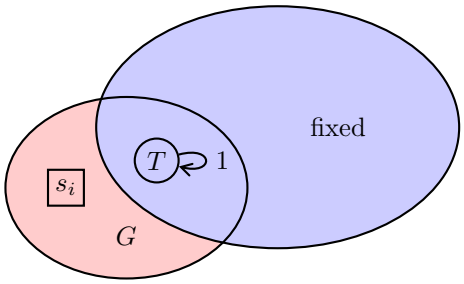


Figure 3 The blue area has been fixed in the first $i-1$ iterations. The smaller ellipse is the set G, where strategy σ does well. The red area will be fixed using σ. Under σ, a run that starts from s_i is likely to lead to the target T without ever leaving G.

We also accomplish goal (B): the difference between $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ is that σ is fixed on G, but σ performs well from G on. So we obtain for *all* states s:

$$\mathtt{val}_{\mathcal{M}_{i+1},s}(\mathtt{Reach}(T)) \;\geq\; \mathtt{val}_{\mathcal{M}_i,s}(\mathtt{Reach}(T)) - \varepsilon_i \tag{3}$$

In slightly more detail, this inequality holds because any strategy in $\mathcal{M}_i$ can be transformed into a strategy in $\mathcal{M}_{i+1}$, with the difference that once the newly fixed part G is entered, the strategy switches to the strategy σ, which (by the definition of $\mathcal{M}_{i+1}$) is consistent with the fixing and (by the definition of G) is ε_i-optimal from there. We give a detailed proof of (3) in Section 5.2. This completes the description of round i.

Let $\varepsilon \in (0,1)$, and for all $i \geq 1$, choose $\varepsilon_i \stackrel{\text{def}}{=} \frac{\varepsilon}{2} \cdot 2^{-i}$. Let σ be an arbitrary MD strategy that is compatible with all fixings. (This strategy σ is actually unique.) It follows that σ is playable in all $\mathcal{M}_i$. Consider an arbitrary state s_i. Then it follows from (3) that the value $\mathtt{val}_{\mathcal{M}_i,s_i}(\mathtt{Reach}(T))$ is not much lower than $\mathtt{val}_{\mathcal{M},s_i}(\mathtt{Reach}(T))$, and from (2) that σ realizes most of this value, implying for all $i \geq 1$:

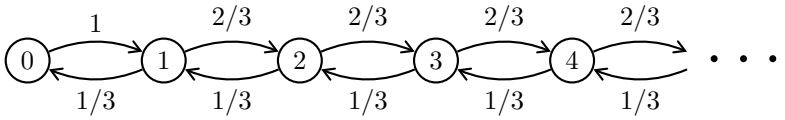
$$\mathbb{P}^{\sigma}_{\mathcal{M},s_i}(\mathtt{Reach}(T)) \;\geq\; \mathtt{val}_{\mathcal{M},s_i}(\mathtt{Reach}(T)) - \varepsilon \tag{4}$$

We give a detailed proof of (4) in Section 5.3. Thus, the MD strategy σ is ε-optimal for all states. ◀

▶ Remark 4. Instead of $\mathtt{Reach}(T)$ the proof above also works for many other objectives, including so-called *tail* events. See Section 3.3 for more discussion about tail events.

3.3 Lévy's Zero-One Law

Consider a Markov chain $\mathcal{M} = (S, \emptyset, S_{\bigcirc}, \longrightarrow, P)$, i.e., an MDP without controlled states. Recall that an event is a set of runs $s_0 s_1 \cdots$. For example, in the Markov chain



we may define an event E as the set of all runs that start in state 0 and revisit state 0 exactly once. Starting in state 1, the probability of ever visiting state 0 can be calculated to be $\frac{1}{2}$. It follows that, starting in state 0, the probability of E is $\frac{1}{2} \cdot (1 - \frac{1}{2}) = \frac{1}{4}$; formally, $\mathbb{P}_0(E) = \frac{1}{4}$. There are two ways of failing to satisfy E: one is to never revisit state 0, the other is to revisit

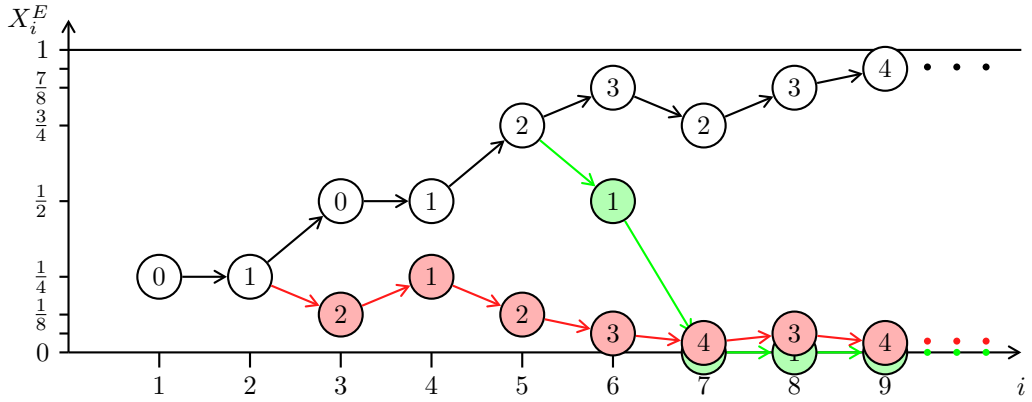


Figure 4 For the event E (exactly one revisit to state 0), three sample runs starting from state 0 are depicted, along with their values $X_1^E, \ldots, X_9^E$. The partially covered "green" run revisits state 0 for a second time, causing $0 = X_7^E = X_8^E = \ldots$ Lévy's zero-one law asserts that X_i^E almost surely converges to 0 or 1.

it more than once. It is because of the latter that for any given finite prefix of a run, we can never be sure that E will hold. For example, if the run starts with 010123, then we have already revisited state 0 and it is unlikely that we will ever do so again: the probability is only $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. So, given this prefix 010123, the probability of satisfying E is $\frac{7}{8}$. However, in this example, no matter what prefix is given, the probability of satisfying E cannot be 1.

Let us define a sequence of random variables, $X_1^E, X_2^E, \ldots$, so that each X_i^E maps a run ρ to the probability that event E will be satisfied, given the prefix of ρ of length i. For example, if $\rho = 010123234\cdots$, then we previously discussed that $X_1^E(\rho) = \frac{1}{4}$ and that $X_6^E(\rho) = \frac{7}{8}$. It is (a consequence of) Lévy's zero-one law that the sequence $X_1^E, X_2^E, \ldots$[1] converges to 0 or 1 almost surely (see also Figure 4):

▶ **Theorem 5** (Lévy's zero-one law for Markov chains). *Let $\mathcal{M} = (S, \emptyset, S_\bigcirc, \longrightarrow, P)$ be a Markov chain, $s_0 \in S$, and E an event of runs starting in s_0. We have $\lim_{i\to\infty} X_i^E \in \{0,1\}$ (and hence this limit exists) almost surely; more formally:*

$$\mathbb{P}_{s_0}\left(\left\{\rho \;\middle|\; \lim_{i\to\infty} X_i^E(\rho) \in \{0,1\}\right\}\right) \;=\; 1\,.$$

Moreover, up to a null set of runs, the limit is 1 for those runs satisfying E, and 0 for those runs not satisfying E.

As a consequence, the probability of E is equal to the probability that the limit is 1.

For *tail* objectives E, Lévy's zero-one law becomes simpler and clearer. A tail objective is an objective whose occurrence is independent of any finite prefix. The objective E from the example above is not a tail objective because the number of revisits to state 0 may change if we cut off or add a finite prefix. For an example of a tail objective, suppose that the states of an arbitrary Markov chain are classified as accepting and non-accepting states. The *Büchi* objective consists of those runs that visit accepting states infinitely often. Büchi is a tail objective: for any run we can cut off or add any finite prefix without changing whether the run satisfies Büchi. We can also view reachability objectives as tail objectives, provided that the target is a sink, which is often a harmless assumption.

[1] It is conventional to write X_i^E for $X_i^E(\rho)$ if the run (such as a random run produced by $\mathcal{M}$) is understood.

For tail objectives E, the random variable X_i^E depends only on the ith visited state (and not also on the first $i-1$ states as in the general case), and for any run $s_1 s_2 s_3 \cdots$ we have $X_i^E = \mathbb{P}_{s_i}(E)$. For a tail objective E and any state s, a picture analogous to Figure 4 would show each occurrence of state s on the same height, independently of the partial run leading up to the occurrence. As a consequence of Lévy's zero-one law, for any tail objective E, the events

$$E \text{ and } \left\{ s_1 s_2 \cdots \;\middle|\; \lim_{i\to\infty} \mathbb{P}_{s_i}(E) = 1 \right\} \text{ are equal up to a null set.} \tag{5}$$

This can be used, in conjunction with Ornstein's Theorem 3 about uniform ε-optimal MD strategies, to construct almost surely winning MD strategies:

▶ **Theorem 6** ([12], uniform almost surely winning MD strategies). *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and consider the reachability objective $\texttt{Reach}(T)$ for some $T \subseteq S$. Let $S_0 \subseteq S$ be the set of states from which there exists an almost surely winning strategy. Then there exists a single MD strategy σ that is almost surely winning for all $s \in S_0$ at the same time; formally, $\mathbb{P}_s^\sigma(\texttt{Reach}(T)) = 1$ for all $s \in S_0$.*

Proof. Obtain from $\mathcal{M}$ an MDP $\mathcal{M}_0$ by restricting the state space to S_0 and eliminating all transitions that leave S_0. In $\mathcal{M}_0$ all states have an almost surely winning strategy, as an almost surely winning strategy may never enter a state that does not have an almost surely winning strategy. By Theorem 3 there exists, for $\mathcal{M}_0$, a uniform $\frac{1}{2}$-optimal MD strategy σ. Then $\mathbb{P}^\sigma_{\mathcal{M}_0,s}(\texttt{Reach}(T)) \geq \frac{1}{2}$ holds for all states. For any run $s_0 s_1 \cdots$ in $\mathcal{M}_0$ we have $\mathbb{P}^\sigma_{\mathcal{M}_0,s_i}(\neg\texttt{Reach}(T)) \leq \frac{1}{2}$ for all i; in particular, the sequence $\left(\mathbb{P}^\sigma_{\mathcal{M}_0,s_i}(\neg\texttt{Reach}(T))\right)_i$ does not converge to 1. Using (5) for $E = \neg\texttt{Reach}(T)$, we obtain for all $s \in S_0$ that $\mathbb{P}_s^\sigma(E) = 0$, hence, $\mathbb{P}_s^\sigma(\texttt{Reach}(T)) = 1$. ◀

3.4 The Flag Construction

We now move from reachability to co-Büchi objectives: here, a subset of states are marked as "bad", and the goal is to visit bad states only finitely often. Co-Büchi is more general than both reachability and safety objectives: for reachability, make the target a sink and mark all other states as bad; for safety, make the states to be avoided a sink and mark them as bad.

In this section we focus on almost surely winning strategies, and we will argue that for co-Büchi objectives almost surely winning strategies can be chosen MD. However, this does not always hold for infinitely branching MDPs, as the example in Figure 5 shows. Therefore, we assume in the rest of the section that the MDP is finitely branching. We will show:

▶ **Theorem 7** ([11]). *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be a finitely branching MDP, and consider a co-Büchi objective $\texttt{co-Büchi}(B)$ for some set $B \subseteq S$ of bad states. Let $S_0 \subseteq S$ be the set of states from which there exists an almost surely winning strategy. Then there exists a single MD strategy σ that is almost surely winning for all $s \in S_0$ at the same time; formally, $\mathbb{P}_s^\sigma(\texttt{co-Büchi}(B)) = 1$ for all $s \in S_0$.*

In order to prove Theorem 7, a *safety strategy* may appear promising: in each state minimize the probability of ever visiting a bad state again. The appeal of a safety strategy is twofold:

- If a safety strategy succeeds in never visiting a bad state again, then clearly it visits bad states only finitely often.
- A safety strategy can be chosen uniformly MD (in finitely branching MDPs): in every controlled state pick the successor with the best value.[2]

[2] Such an approach for constructing an optimal strategy does not work for reachability or more general objectives: intuitively, this approach cannot guarantee "progress" towards the goal.

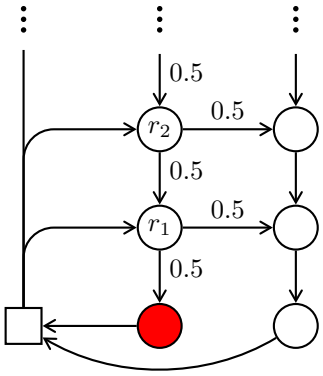


Figure 5 In infinitely branching MDPs with co-Büchi objectives, almost surely winning strategies cannot always be chosen MD. In the MDP above, the bad state is marked red. There exists an almost surely winning strategy, but it requires memory in order to choose r_i for ever higher i.

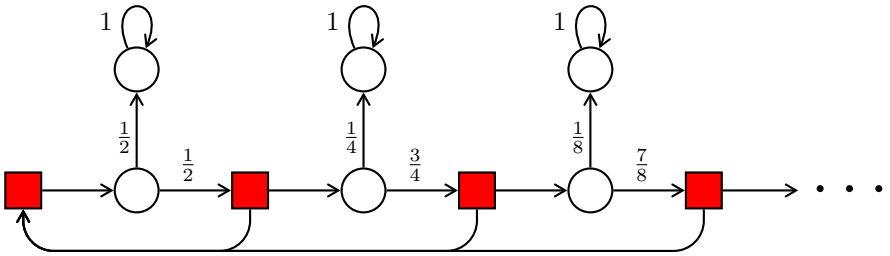


Figure 6 In this MDP (with bad states marked red), a safety strategy always chooses the right outgoing transition, chasing after the ever smaller chance of entering a safe sink state without re-entering any bad state. Starting from the leftmost state, the probability that this strategy achieves the co-Büchi objective is less than 0.72. The "opposite" strategy, which always returns to the leftmost state, succeeds almost surely.

But a safety strategy alone does not suffice for co-Büchi, as the example in Figure 6 shows. In the following proof we construct an almost surely winning MD strategy for co-Büchi by combining an MD strategy for safety with an MD strategy for reachability.

Proof sketch of Theorem 7. Similarly as in the proof of Theorem 6, we can assume without loss of generality that for all states there exists an almost surely winning strategy. We will show that there exists a single MD strategy that is almost surely winning for all states.

We have mentioned previously that there exists an MD uniformly optimal safety strategy σ_{safe}, i.e., for each state, σ_{safe} minimizes the probability of ever revisiting a bad state. For $x \in [0,1]$ define $Safe(x) \subseteq S$ as the set of states with *safety level* at least x. By safety level we mean the probability of never visiting another bad state, assuming σ_{safe} is played. See Figure 7 for an abstract visualization. We *fix* σ_{safe} in $Safe(\frac{1}{3})$, i.e., in the following we will only consider strategies that are compatible with σ_{safe} in $Safe(\frac{1}{3})$, see the right side of Figure 7.

After this fixing, every state still has an almost surely winning strategy. Indeed, consider any state and its almost surely winning strategy before the fixing. We modify the strategy as follows. First we play it as before, but if and when we reach $Safe(\frac{1}{3})$, we switch to σ_{safe}. Now the probability is at least $\frac{1}{3}$ that we never visit a bad state again and thus also achieve the co-Büchi objective. If we do visit a bad state again, we revert to a strategy that is almost surely winning from that bad state in the original MDP. We follow that strategy until we

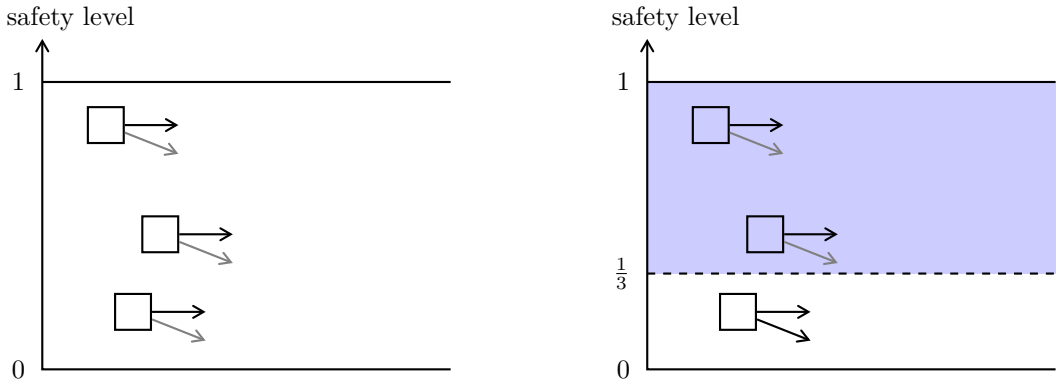


Figure 7 Left: In this diagram the controlled states are arranged according to their safety level. A safety strategy entails not to pick successor states with smaller safety level, so grey transitions are not used. Right: The optimal safety strategy σ_{safe} is fixed for the states with safety level at least $\frac{1}{3}$.

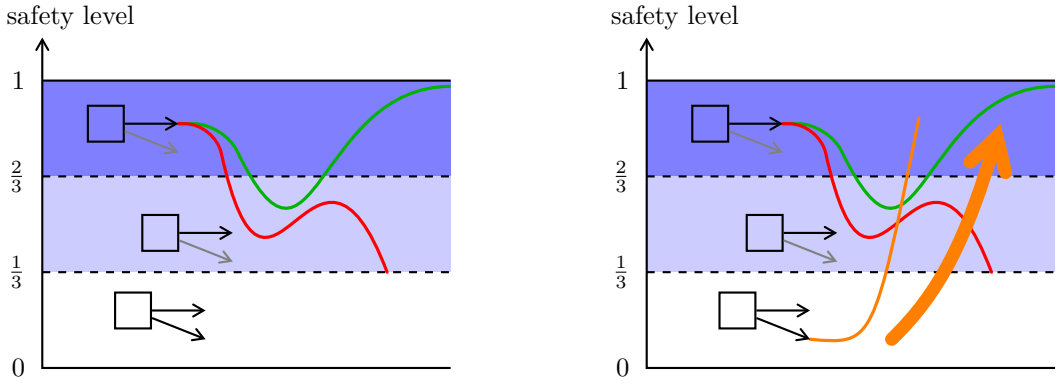


Figure 8 Left: Starting from a state with safety level at least $\frac{2}{3}$ the probability of keeping a safety level of at least $\frac{1}{3}$ forever is at least $\frac{1}{2}$. Right: For the states with safety level less than $\frac{1}{3}$ we aim at reaching safety level at least $\frac{2}{3}$.

possibly reach $Safe(\frac{1}{3})$ again. At this point we switch again to σ_{safe}, thus forever avoiding the bad states with a fresh chance of at least $\frac{1}{3}$. Continuing in this way, we win almost surely. Note that this strategy is not MD, as we have to remember in which phase we are: at any point we either follow a winning strategy of the original MDP, or follow σ_{safe}. We have merely argued here that having fixed σ_{safe} in $Safe(\frac{1}{3})$ has not done any harm.

Before we define an MD strategy for the rest of the state space, consider a state $s \in Safe(\frac{2}{3})$, i.e., s has an even higher safety level of at least $\frac{2}{3}$. Since $Safe(\frac{2}{3}) \subseteq Safe(\frac{1}{3})$, the MD strategy σ_{safe} has been fixed there. Two things might happen starting at s (see the left side of Figure 8): either the run remains in $Safe(\frac{1}{3})$ forever and never visits a bad state; or it eventually leaves $Safe(\frac{1}{3})$ (or even visits a bad state). The second case (leaving $Safe(\frac{1}{3})$ or visiting a bad state) cannot have a very large probability: after all, we start with safety level at least $\frac{2}{3}$, so if the safety level is very likely to drop below $\frac{1}{3}$, we were not very safe to start with. Doing the maths shows that the probability of the second case is at most $\frac{1}{2}$. So starting from $Safe(\frac{2}{3})$, the probability of avoiding bad states forever is at least $\frac{1}{2}$, no matter what strategy is played outside of $Safe(\frac{1}{3})$.

We still need to define an MD strategy for the part of the state space with safety level less than $\frac{1}{3}$. Our ambition is to define it so that from every state we reach almost surely $Safe(\frac{2}{3})$, see the right side of Figure 8. If we succeed in this goal, then we achieve the co-Büchi objective almost surely, because every time we reach $Safe(\frac{2}{3})$ we receive a fresh chance of $\frac{1}{2}$ to avoid bad states forever, as just argued.

First we argue that for each state there is *some* strategy to reach $Safe(\frac{2}{3})$ almost surely. Recall that we have shown above that after the fixing in $Safe(\frac{1}{3})$ every state s still has a strategy σ_s to achieve the co-Büchi objective almost surely. We argue that σ_s reaches $Safe(\frac{2}{3})$ almost surely. Indeed, whenever a run is outside of $Safe(\frac{2}{3})$ there is a risk of more than $\frac{1}{3}$ to visit a bad state. It follows that there is a risk of at least $\frac{1}{3}$ to visit a bad state within a finite time horizon. Since σ_s almost surely achieves the co-Büchi objective, it must avoid that this risk materializes infinitely often. Hence σ_s almost surely reaches $Safe(\frac{2}{3})$.

Using Theorem 6 it follows that, in the MDP after having fixed σ_{safe} in $Safe(\frac{1}{3})$, there is uniform almost surely winning MD strategy σ_{reach} for $\texttt{Reach}(Safe(\frac{2}{3}))$. In summary, here is our almost surely winning MD strategy for co-Büchi: in $Safe(\frac{1}{3})$ play σ_{safe}, and elsewhere play σ_{reach}. The key point is that these two MD strategies are not conflicting. ◀

▶ Remark 8. Generalizations of Theorem 7 are also considered in [11]:

1. A similar proof shows a version of Theorem 7 for ε-optimal strategies.
2. Theorem 7 generalizes to $\{0,1,2\}$-parity objectives, which also encompass Büchi objectives. The theorem further generalizes from almost surely winning to *optimal* strategies as follows: The set $S_0 \subseteq S$ can be taken as the set of states from which there exists an optimal strategy (note that an almost surely winning strategy is optimal). Then there exists a single MD strategy that is optimal for all states in S_0.
3. Theorem 7 does not hold for $\{1,2,3\}$-parity objectives.

4 Markov Strategies and Generalizations

We describe a technique to prove the existence of ε-optimal Markov strategies (resp. Markov strategies with one extra bit of memory) for certain types of objectives, based on the work on Büchi objectives in [10].

Obtaining Markov strategies via acyclic MDPs. *Markov strategies* are strategies that base their decisions only on the current state and the number of steps in the history of the run from some initial state s_0. Thus they do use infinite memory, but only in a very restricted form by maintaining an unbounded step counter. Slightly more general are *1-bit Markov strategies* that use one extra bit of extra memory in addition to a step counter.

The existence of ε-optimal (1-bit) Markov strategies for some objective φ on countable MDPs can be proven by first studying the strategy complexity of φ on *acyclic* MDPs, i.e., MDPs where the underlying transition graph is a directed acyclic graph (DAG). Note that a DAG is more general than a tree. If the transition graph is a tree with root s_0 then there always exist ε-optimal positional strategies for any objective, since the entire history is implicit in the current state. This does not hold for a DAG, since the same state s could be reached via (possibly infinitely many) different paths from s_0.

However, for every countable MDP $\mathcal{M}$ with initial state s_0, there is a corresponding acyclic MDP $\mathcal{M}'$ that encodes the step counter into the states, i.e., the states of $\mathcal{M}'$ are of the form (s,i) where s is a state of $\mathcal{M}$ and $i \in \mathbb{N}$ counts the number of steps. Then for every ε-optimal positional strategy for φ in $\mathcal{M}'$ there is a corresponding ε-optimal Markov strategy for φ in $\mathcal{M}$, and vice-versa [10]. Thus, if ε-optimal positional strategies for φ exist in *all*

acyclic MDPs then ε-optimal Markov strategies for φ exist in general countable MDPs. The reverse implication does not hold, however, since not all acyclic MDPs encode a step counter, e.g., if some state has infinite in-degree and can be reached from the initial state via paths of arbitrary length.

Acyclic and finitely branching MDPs have nice properties that make it easier to infer the existence of simpler ε-optimal strategies. In the following, we first observe the behavior of a general ε-optimal strategy, and then show how a 1-bit strategy can closely match it (for certain types of objectives).

Observing the behavior of an ε-optimal strategy. Consider an acyclic and finitely branching MDP with initial state s_0. (One could also have a finite set of initial states as in [10], but we use a single state to simplify the presentation.) Let σ be an arbitrary ε-optimal strategy from the initial state s_0. We now observe its behavior, i.e., the induced runs. Let $\mathsf{bubble}_k(\{s_0\})$ be the set of states that can be reached from the initial state s_0 within at most k steps. This set is finite for every finite k, since our MDP is finitely branching. Note that, by acyclicity, any given run can visit a given finite set of states X only finitely often (at most $|X|$ times). However, this does not imply that the probability of re-visiting X must eventually become zero. E.g., it is possible that in the i-th state of some run the probability of re-visiting X (in continuations of this run) is 2^{-i} (i.e., re-visiting X remains always possible, but does not happen almost surely).

Still, a weaker property does hold in acyclic and finitely branching MDPs. It follows from acyclicity [10, Lemma 10] that, after a sufficiently large number of steps, runs are arbitrarily unlikely to visit any given finite set of states again. In particular this holds for the finite set $\mathsf{bubble}_k(\{s_0\})$.

Formally, for every k and $\delta > 0$ there exists some l such that the probability of visiting $\mathsf{bubble}_k(\{s_0\})$ after step l is $\leq \delta$. By definition, states *outside* the set $\mathsf{bubble}_l(\{s_0\})$ are reachable only after a number of steps that is strictly larger than l. Therefore, it is unlikely (probability $\leq \delta$) to visit $\mathsf{bubble}_k(\{s_0\})$ again after some state $s \notin \mathsf{bubble}_l(\{s_0\})$ has been visited for the first time.

These observations allow to define a decreasing sequence $\delta_i \stackrel{\text{def}}{=} \varepsilon \cdot 2^{-i}$ of small errors and sufficiently large and increasing numbers k_i and l_i with $k_i < l_i < k_{i+1}$ for $i \geq 1$ such that for the finite sets $K_i \stackrel{\text{def}}{=} \mathsf{bubble}_{k_i}(\{s_0\})$ and $L_i \stackrel{\text{def}}{=} \mathsf{bubble}_{l_i}(\{s_0\})$ it is unlikely (probability $\leq \delta_i$) to visit K_i after leaving L_i (for the first time). I.e., runs like π_2 in Figure 9 are unlikely. However, the probability of leaving K_i and later returning to K_i (even multiple times) before leaving L_i may be large.

Note that we still have a lot of freedom to choose the numbers k_i and l_i. For the numbers k_i we just need $l_i < k_{i+1}$. The minimal required size of l_i depends on k_i, σ and δ_i, but l_i can be chosen arbitrarily larger than this minimal size.

Let now SEQ be the objective of never visiting a set K_i after leaving L_i (for the first time) for any $i \geq 1$. (I.e., SEQ depends on the chosen numbers k_i, l_i.)

The strategy σ is not only ε-optimal for the objective φ from state s_0, but also 2ε-optimal for the stronger objective $\varphi \wedge \mathsf{SEQ}$, since $\sum_{i \geq 1} \delta_i \leq \varepsilon$.

Constructing simpler strategies. For certain types of objectives φ (e.g., Büchi objectives), one can exploit this pattern of SEQ to construct simpler (1-bit) ε-optimal strategies for φ in acyclic and finitely branching MDPs. This then yields ε-optimal 1-bit Markov strategies in general finitely branching MDPs.

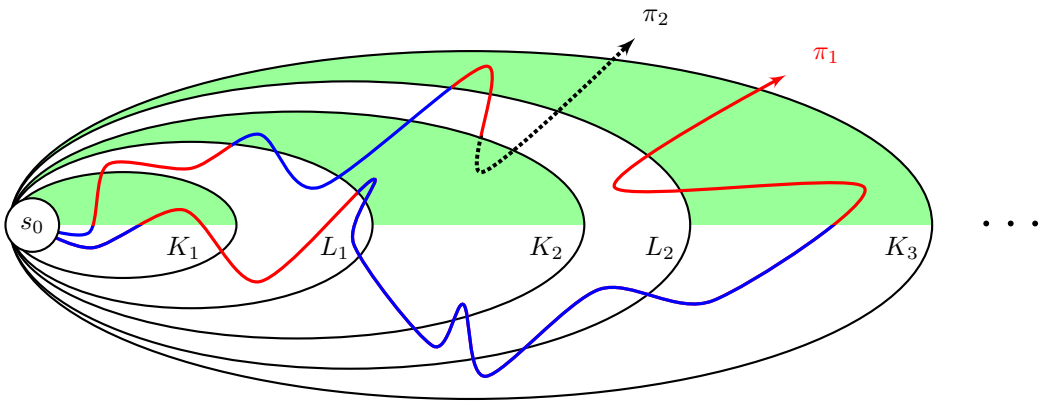


Figure 9 Updates of the extra mode bit along runs π_1, π_2, drawn in blue while the memory-bit is one and in red while the bit is zero. The run π_2 violates SEQ and is drawn as a dotted line once it does. Upon entering the green zone of $K_i \setminus K_{i-1}$, the runs attain the local objective φ_i and flip the mode bit.

Suppose that $\varphi \wedge \mathsf{SEQ}$ can be decomposed into an infinite sequence of local sub-objectives $\varphi_i \wedge \mathsf{SEQ}$ such that, with arbitrarily high probability, satisfying $\varphi_i \wedge \mathsf{SEQ}$ in each finite set $K_i \setminus K_{i-1}$ implies $\varphi \wedge \mathsf{SEQ}$ overall, and vice-versa. (E.g., if φ is a Büchi objective to visit a given set of states F infinitely often then φ_i is the objective to visit the subset of F inside $K_i \setminus K_{i-1}$ (i.e., to visit $F \cap (K_i \setminus K_{i-1})$); cf. [10].) Note that the "vice-versa" part often depends on the fact that the numbers k_i can be chosen sufficiently large to get a sufficiently high probability of satisfying φ_i inside $K_i \setminus K_{i-1}$.

Of course, not every objective φ can be decomposed in this way, e.g., in parity objectives, different runs can win by different colors and local conditions φ_i are insufficient.

Now suppose that in the MDP induced by the finite subspace K_i there exist ε-optimal positional strategies σ_i that attain a high probability of φ_i in $K_i \setminus K_{i-1}$, and additionally maintain a high value w.r.t. future objectives φ_j in $K_j \setminus K_{j-1}$ for all $j > i$. I.e., σ_i has a high attainment for the local sub-objective without compromising future sub-objectives.

One extra bit. The above suggests a scheme to construct an ε-optimal positional strategy σ' for $\varphi \wedge \mathsf{SEQ}$ by playing each local positional strategy σ_i inside $K_i \setminus K_{i-1}$.

However, this is not always sufficient. The problem is that, when playing in $K_i \setminus K_{i-1}$, a run might temporarily go back into the set K_{i-1} (though not into the states that this particular run has previously visited, due to acyclicity). If this run has never yet left L_{i-1}, then going back to K_{i-1} is allowed by SEQ and can be necessary (or even unavoidable). (In contrast, once one has left L_{i-1}, it is possible to henceforth avoid K_{i-1} and still attain φ with high probability, as witnessed by the strategy σ.) But back in K_{i-1} the strategy σ' would play σ_{i-1} towards objective φ_{i-1} (that had already been attained previously) instead of focusing on the current objective φ_i. Although the run will inevitably (by acyclicity) exit K_{i-1} again, it might re-visit K_{i-1} many times, and thus switch the focus back to φ_{i-1} many times. I.e., the strategy σ' might attempt to re-attain the previous objective φ_{i-1} many times over, instead of permanently switching the focus to φ_i once φ_{i-1} has been attained once. Switching the focus back to the previous objective φ_{i-1} too often is wasteful and might damage the ability to attain future objectives φ_j for the $j \geq i$ with high probability, e.g., due to a trade-off between current and future objectives. So this strategy σ' might not always succeed for objective φ (e.g., it cannot work for Büchi objectives [10]).

Instead we want the strategy to always focus on the next objective φ_i after it has completed the previous objective φ_{i-1}. To this end, we need one extra bit of memory, called the *mode bit*, to distinguish two modes of playing: current-mode and next-mode. The mode bit is used to remember whether one has already attained φ_i in $K_i \setminus K_{i-1}$. Upon attaining φ_i in $K_i \setminus K_{i-1}$, one switches from current-mode to next-mode.

One interprets the content of the mode bit (0 or 1) differently for even and odd numbers i, so that the next-mode for $K_i \setminus K_{i-1}$ is the current-mode for $K_{i+1} \setminus K_i$ (and vice-versa). This means that if the strategy plays in next-mode in $K_i \setminus K_{i-1}$ then upon entering $K_{i+1} \setminus K_i$ it automatically switches to current-mode. Dually, if it plays in current-mode in $K_{i+1} \setminus K_i$ and temporarily goes back into $K_i \setminus K_{i-1}$ then it automatically switches to next-mode; cf. Figure 9. The strategy pursues different local goals, depending on the mode bit.

- In current-mode, it continues to focus on attaining φ_i in K_i. Temporarily going back to K_{i-1} does not change the focus on φ_i, because current-mode for K_i is next-mode for K_{i-1}. By suitably choosing φ_i and k_i, one can ensure with high probability that φ_i is attained only in $K_i \setminus L_{i-1}$, i.e., *after* leaving L_{i-1}. So, since one has already left L_{i-1} before this success, it is possible with high probability to avoid K_{i-1} afterwards. In particular, after attaining φ_i in $K_i \setminus L_{i-1}$ the strategy switches the mode-bit to next-mode. The value of the mode bit is the same for next-mode in K_i and for current-mode in K_{i-1}. However, there is only a small danger of ever confusing these, since the probability of visiting K_{i-1} after leaving L_{i-1} is small.
- In next-mode, the focus is on leaving K_i to reach $K_{i+1} \setminus K_i$ and attaining the next objective φ_{i+1}. In particular, upon entering $K_{i+1} \setminus K_i$ the mode bit is interpreted as current mode for $K_{i+1} \setminus K_i$. Moreover, it is then not a problem if the run temporarily goes back from $L_i \setminus K_i$ into K_i, because the focus remains on φ_{i+1} (since current-mode in K_{i+1} is next-mode in K_i).

By combining the positional strategies for the local objectives φ_i with the extra mode bit, one obtains ε-optimal 1-bit strategies on all acyclic finitely branching MDPs. This yields 1-bit Markov strategies on general finitely branching MDPs by the argument above.

For Büchi objectives, one can encode infinite branching into finite branching by a gadget similar to the one used in the proof of Lemma 2. Moreover, the local strategies σ_i can be chosen MD. Thus one obtains deterministic 1-bit Markov ε-optimal strategies. There is also a matching lower bound.

▶ **Theorem 9** ([10], ε-optimal deterministic 1-bit Markov strategies for Büchi objectives)**.** *Given a countable MDP and a Büchi objective, for every $\varepsilon > 0$ and initial state s_0, there exists an ε-optimal deterministic 1-bit Markov strategy. Moreover, neither randomized Markov strategies nor randomized finite-memory strategies are sufficient.*

▶ Remark 10. The whole argument can, of course, be generalized. If the strategies for the local objectives φ_i in acyclic MDPs are not positional but use finite memory, say m bits, then one obtains $(m+1)$-bit Markov strategies in general MDPs.

5 Missing Proof Details

5.1 Proof of Equation (1)

Proof. For a state $s \in S \setminus G$, define the event L_s as the set of runs that leave G such that s is the first visited state in $S \setminus G$. Then we have:

$$\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\mathtt{Reach}(S \setminus G)) \;=\; \sum_{s \in S \setminus G} \mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(L_s) \tag{6}$$

Since T is a sink and using the Markov property:

$$\begin{aligned}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) \;=\; & \mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\neg\texttt{Reach}(S\setminus G)\wedge\texttt{Reach}(T)) + \\ & \sum_{s\in S\setminus G}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(L_s)\cdot\mathbb{P}^{\sigma}_{\mathcal{M}_i,s}(\texttt{Reach}(T))\end{aligned} \tag{7}$$

By the definition of G it follows:

$$\begin{aligned}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) \;\leq\; & \mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\neg\texttt{Reach}(S\setminus G)\wedge\texttt{Reach}(T)) + \\ & \sum_{s\in S\setminus G}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(L_s)\cdot(\texttt{val}_{\mathcal{M}_i,s}(\texttt{Reach}(T))-\varepsilon_i)\end{aligned} \tag{8}$$

On the other hand, σ is ε_i^2-optimal for s_i, hence:

$$\begin{aligned}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) \;&\geq\; -\varepsilon_i^2+\texttt{val}_{\mathcal{M}_i,s}(\texttt{Reach}(T)) \\ &\geq\; -\varepsilon_i^2+\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)\wedge\neg\texttt{Reach}(S\setminus G)) + \\ &\qquad \sum_{s\in S\setminus G}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(L_s)\cdot\texttt{val}_{\mathcal{M}_i,s}(\texttt{Reach}(T))\end{aligned} \tag{9}$$

By combining (8) and (9) we obtain:

$$\varepsilon_i^2 \;\geq\; \varepsilon_i\cdot\sum_{s\in S\setminus G}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(L_s) \;=\; \varepsilon_i\cdot\mathbb{P}^{\sigma}_{\mathcal{M}_i,s_i}(\texttt{Reach}(S\setminus G)) \qquad \blacktriangleleft$$

5.2 Proof of Equation (3)

Proof. For a state $s'\in G$, define the event $E_{s'}$ as the set of runs that enter G such that s' is the first visited state in G. Fix any state $s\in S$ and any strategy σ_i in $\mathcal{M}_i$. We transform σ_i into a strategy σ_{i+1} in $\mathcal{M}_{i+1}$ such that σ_{i+1} behaves like σ_i until G is entered, at which point σ_{i+1} switches to the MD strategy σ, which we recall is compatible with $\mathcal{M}_{i+1}$ and is ε_i-optimal from G in $\mathcal{M}_i$. To show (3) it suffices to show that $\mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s}(\texttt{Reach}(T))\geq\mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(\texttt{Reach}(T))-\varepsilon_i$. We have:

$$\begin{aligned}\mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s}(\texttt{Reach}(T)) \;=\; & \mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s}(\neg\texttt{Reach}(G)\wedge\texttt{Reach}(T)) + && \text{as } T \text{ is a sink} \\ & \sum_{s'\in G}\mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s}(E_{s'})\cdot\mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s'}(\texttt{Reach}(T)) && \text{Markov property} \\ =\; & \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(\neg\texttt{Reach}(G)\wedge\texttt{Reach}(T)) + && \text{using def. of } \sigma_{i+1} \\ & \sum_{s'\in G}\mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(E_{s'})\cdot\mathbb{P}^{\sigma}_{\mathcal{M}_i,s'}(\texttt{Reach}(T)) &&\end{aligned}$$

Further we have for all $s'\in G$:

$$\begin{aligned}\mathbb{P}^{\sigma}_{\mathcal{M}_i,s'}(\texttt{Reach}(T)) \;&\geq\; \texttt{val}_{\mathcal{M}_i,s'}(\texttt{Reach}(T))-\varepsilon_i && \text{as } s'\in G \\ &\geq\; \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s'}(\texttt{Reach}(T))-\varepsilon_i &&\end{aligned}$$

Plugging this in above, we obtain:

$$\begin{aligned}
\mathbb{P}^{\sigma_{i+1}}_{\mathcal{M}_{i+1},s}(\texttt{Reach}(T)) &\geq \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(\neg\texttt{Reach}(G) \wedge \texttt{Reach}(T)) + \\
&\quad \sum_{s'\in G} \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(E_{s'}) \cdot (\mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s'}(\texttt{Reach}(T)) - \varepsilon_i) \\
&\geq \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(\neg\texttt{Reach}(G) \wedge \texttt{Reach}(T)) + \\
&\quad \Big(\sum_{s'\in G} \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(E_{s'}) \cdot \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s'}(\texttt{Reach}(T))\Big) - \varepsilon_i \\
&= \mathbb{P}^{\sigma_i}_{\mathcal{M}_i,s}(\texttt{Reach}(T)) - \varepsilon_i \qquad \blacktriangleleft
\end{aligned}$$

5.3 Proof of Equation (4)

Proof. For all $i \geq 1$ we have:

$$\begin{aligned}
\mathbb{P}^{\sigma}_{\mathcal{M},s_i}(\texttt{Reach}(T)) &\geq \texttt{val}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) - \varepsilon_i - \varepsilon_i^2 && \text{by (2)} \\
&\geq \texttt{val}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) - 2\varepsilon_i && \text{as } \varepsilon_i < 1 \\
&\geq \texttt{val}_{\mathcal{M}_i,s_i}(\texttt{Reach}(T)) - \frac{\varepsilon}{2} && \text{choice of } \varepsilon_i \\
&\geq \texttt{val}_{\mathcal{M},s_i}(\texttt{Reach}(T)) - \sum_{j=1}^{i-1}\varepsilon_j - \frac{\varepsilon}{2} && \text{by (3)} \\
&\geq \texttt{val}_{\mathcal{M},s_i}(\texttt{Reach}(T)) - \varepsilon && \text{choice of } \varepsilon_j \qquad \blacktriangleleft
\end{aligned}$$

References

1 P. Abbeel and A. Y. Ng. Learning first-order Markov models for control. In *Advances in Neural Information Processing Systems 17*, pages 1–8. MIT Press, 2004. URL: `http://papers.nips.cc/paper/2569-learning-first-order-markov-models-for-control`.

2 C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.

3 V. D. Blondel and J. N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

4 N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance*. Springer-Verlag Berlin Heidelberg, 2011.

5 K. Chatterjee, L. de Alfaro, and T. Henzinger. Trading memory for randomness. In *Annual Conference on Quantitative Evaluation of Systems*, pages 206–217. IEEE Computer Society Press, 2004.

6 K. Chatterjee and T. Henzinger. A survey of stochastic ω-regular games. *Journal of Computer and System Sciences*, 78(2):394–413, 2012.

7 K. Chatterjee, M. Jurdziński, and T. Henzinger. Quantitative stochastic parity games. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 121–130, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=982792.982808`.

8 E. M. Clarke, T. A. Henzinger, H. Veith, and R. Bloem, editors. *Handbook of Model Checking*. Springer, 2018. `doi:10.1007/978-3-319-10575-8`.

9 W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley & Sons, second edition, 1966.

10 S. Kiefer, R. Mayr, M. Shirmohammadi, and P. Totzke. Büchi objectives in countable MDPs. In *ICALP 2019*, volume 132. LIPIcs, 2019. `doi:10.4230/LIPIcs.ICALP.2019.119`.

11 S. Kiefer, R. Mayr, M. Shirmohammadi, and D. Wojtczak. Parity Objectives in Countable MDPs. In *Annual IEEE Symposium on Logic in Computer Science*, 2017.

12 D. Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 20:563–569, 1969.

13 M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

14 M. Schäl. Markov decision processes in finance and dynamic options. In *Handbook of Markov Decision Processes*, pages 461–487. Springer, 2002.

15 O. Sigaud and O. Buffet. *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons, 2013.

16 R.S. Sutton and A.G Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2018.

17 W. Zielonka. Perfect-information stochastic parity games. In *Foundations of Software Science and Computation Structures*, pages 499–513. Springer, 2004.